基于 slope one 算法改进评分矩阵填充的协同过滤算法研究 *

向小东，邱梓咸

(福州大学 经济与管理学院，福州 350116)

摘　要：为解决协同过滤算法中的数据稀疏性问题，提出了一种改进的协同过滤算法。该算法使用 slope one 算法计算出来的评分预测值来填充评分矩阵中的未评分项目，然后在填充后的用户-项目评分矩阵上通过基于用户的协同过滤方法给出推荐。利用 slope one 算法计算出来的评分预测值作为回填值，既能降低评分矩阵的稀疏性，也保证了回填值的多样性，从而减少均值、中值等单一填充值造成的推荐误差。在 MovieLens-1M 数据集上对本文改进算法和协同过滤算法及均值中心化处理的算法作五折交叉实验，结果表明，基于评分预测值填充数据后的协同过滤算法有效的缓解了数据稀疏性问题，并且有更好的推荐效果。

关键词：slope one 算法；数据稀疏性；协同过滤；数据稀疏性；矩阵填充；电影推荐

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.12.0751

Research on collaborative filtering algorithm based on slope one algorithm to improve score matrix filling

Xiang Xiaodong, Qiu Zixian

(*School of Economics & Management*, *Fuzhou University*, *Fuzhou* 350116, *China*)

Abstract: In order to solve the problem of data sparsity in the collaborative filtering algorithm, this paper proposes an improved collaborative filtering algorithm. The algorithm fills the unrated items in the scoring matrix using the prediction value calculated by the Slope One algorithm and then gives recommendations based on the user-based collaborative filtering method based on the filled user-item scoring matrix. Using the predictive value of Slope One algorithm as the backfill value can not only reduce the sparsity of the scoring matrix, but also ensure the diversity of backfill values, so as to reduce the recommended error caused by the single fill value such as mean value and median value. Half off cross-validation experiments were performed on the movielens-1M dataset. The results show that the collaborative filtering algorithm based on the score prediction data effectively mitigates data sparsity and has better performanceRecommended effect.

Key words: slope one algorithm; data sparsity; personalized recommendation; collaborative filtering; matrix completion; movie recommendation

0　引言

近年来，随着科技的进步和网络技术的不断发展，网络信息以爆炸式增长，导致用户的真正需求淹没在大量无关信息和多样化的产品中，使得用户花费大量的时间还不一定能找到自己需要的商品，信息过载问题亟待有效解决；另一方面，随着社会的发展和人们对个性化的追求，如何满足每个人独特的兴趣嗜好也是个难题。在此背景下，个性化推荐技术应运而生并在影视、新闻、电商平台、高校图书馆等迅速兴起并不断发展。

个性化推荐技术中运用的推荐算法虽不尽相同，但基于协同过滤算法的推荐由于其简单性、准确性、有效性等优势成为目前应用最广泛的个性化推荐算法[1]。数据稀疏性一直以来都是协同过滤算法的一大难题，对此，国内外学者进行了大量的研究[2]。冷亚军等人[3]认为数据稀疏性问题会从近邻搜寻不够准确和近邻评分过少两方面对协同过滤产生不利影响。孟详武等人[4]针对推荐系统中的数据稀疏性和冷启动等问题，对社会化推荐系统在信任推理以及推荐关键技术等方面做了比较全面的综述。孔欣欣等人[5]提出了一种标签权重的评分方法来最大化地降低客观因素对用户评分的影响，有效缓解了用户的评分偏差问题。Li 等人[6]为解决推荐系统数据稀疏性及精度过低的问题，提出了一种结合用户评分信任和用户偏好信任改善评分相似性来计算的协同过滤算法。Songjie 等人[7]指出可以用聚类来改善相似度的计算。Lu 等人[8]提出了一种基于缺失值迭代预测填充的协同过滤算，不仅能够降低数据稀疏性的还提高了用

收稿日期：2017-12-01；**修回日期：**2018-01-29　　**基金项目：**福建省软科学项目（2017R0055）

作者简介：向小东（1973-），男，四川广安人，教授，博士，主要研究方向为管理科学与工程（1467288927@qq.com）；邱梓咸（1993-）硕士研究生，主要研究方向为数据挖掘、个性化推荐、协同过滤.

户相似度计算精度问题。Vozalis 等人[9]把奇异值分解方法和基于项目的方法融合到协同过滤中，降低用户项目矩阵的规模从而有效缓解了数据稀疏的问题。Yang 等人[10]提出一种结合用户信任关系的改进协同过滤算法，通过集成用户提供的稀疏评估数据和稀疏的社交信任网络来提高协作过滤的推荐性能，并在四种数据集上验证了改进算法的推荐效果。

从这些算法的仿真结果可以看出推荐系统的质量得到了提升，但仍然难以完美解决协同过滤算法中存在的所有问题，比如空值填补法中若采用缺省值（即用户评分中值、均值、默认值 0）来回填数据，由于用户的未评分项不可能完全相同，导致信任度不高。本文就近邻评分数据过少，先通过原始数据得到初步的用户相似度和每个用户的近邻，利用 slope one 算法计算评分预测值来填充数据，并基于填充后的数据修正相似度和优化近邻选取集合，最终给出目标用户的推荐列表。实验数据集来自 movielen-1m 的数据集，采用五折交叉实验法进行实证研究，同协同过滤算法和基于均值填充后的协同过滤算法结果进行比较分析。

1　算法描述及模型设计

1.1　算法相关描述

Slope One 算法是一种经典的基于用户-项目评分矩阵的推荐算法，该算法是一个增量算法，对评分较少的用户也可以产生推荐，同时准确度比传统的基于用户和项目的协同过滤算法要好[11]，故本文采用 Slope One 算法来计算预测评分值。与其他类似推荐算法相比，它的最大优点在于算法很简单，易于实现，执行效率高，同时推荐的准确性相对较高[12-14]。

1.1.1 Slope One 算法

Slope One 算法是基于不同项目之间的评分差的线性算法，预测用户对项目评分的个性化算法。主要分两步：

a)计算项目之间的评分差的均值，记为物品间的评分偏差(两项目同时被评分) Dev_{ij} 如式(1)所示。

$$Dev_{ij}=\frac{\sum_{u\in N(i)\cap N(j)}(r_{ui}-r_{uj})}{|N(i)\cap N(j)|} \tag{1}$$

其中：r_{ui}、r_{uj} 分别为用户 u 对项目 i 和项目 j 的评分，$N(i)$、$N(j)$ 分别为对项目 i、j 评过分的用户集合，$N(i)\cap N(j)$ 是对项目 i 和项目 j 都评过分的用户集合，$|N(i)\cap N(j)|$ 是物品 i 和物品 j 都评过分的用户数。

b)根据项目间的评分偏差和用户的历史评分，预测用户对未评分的项目的评分 p_{ui} 如式(2)所示。

$$p_{ui}=\frac{\sum_{i\in N(u)}|N(i)\cap N(j)|\times(r_{uj}+dev_{ij})}{\sum_{i\in N(u)}|N(i)\cap N(j)|} \tag{2}$$

其中：$N(u)$ 是用户 u 评过分的物品。

1.1.2 协同过滤

“协同过滤”一词最早是由 GlodBerg 等人[15]在 20 世纪 90 年代中期开发推荐系统 Tapestry 时提出。近年来随着学者们对协同过滤的深入研究和应用。将各种相关技术同协同过滤推荐算法结合使用，当前协同过滤的一种经典分类方法如图 1 所示。而本文研究的主要内容是对基于内存下的 UBCF 算法的改进。其中，基于内存是在内存当中都需要维护一个庞大的相似度矩阵，并采用一定的启发式为用户推荐，也称为基于领域（domain-based）或基于记忆（memory-based）；基于项目或基于物品（item-based collaborative filtering，IBCF）；基于用户（user-based collaborative filtering，UBCF）。

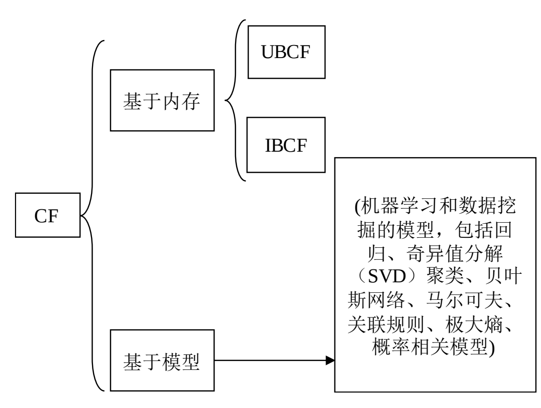


图 1　协同过滤推荐算法分类图

1.1.3 相似度

相似度是指两个用户对同一个项目的喜好程度（用户相似度）或者两个项目被同一个用户喜好的相似程度（项目相似度）。协同过滤算法中常用相似度有：欧式距离相似度、Jaccard 相似度、余弦相似度、修正的余弦相似度、Pearson 相似度[16,17]等。由于余弦相似度在数据越稀疏的情况下，效果更好，故本文在处理初始评分矩阵和填充后的评分矩阵过程中，统一使用余弦相似度。余弦相似度公式如式(3)所示。

$$sim_{(u,v)}=\frac{\sum_{i\in I}R_{u,i}\times R_{v,i}}{\sqrt{\sum_{i\in I}R_{u,i}^2}\sqrt{\sum_{i\in I}R_{u,j}^2}} \tag{3}$$

其中：u、v 代表用户，i 代表项目，i 表示项目集合，$R_{u,i}$ 表示用户 u 对项目 i 的评分值，$R_{u,j}$ 为用户对 u 对项目 j 的评分。

1.1.4 协同过滤一般流程

传统协同过滤算法的推荐流程一般分为四个步骤如图 2 所示，本文改进的工作主要在二三两个步骤之间，使用 slope-one 算法来作空缺值填充，从而构建稀疏度较低的评分矩阵。

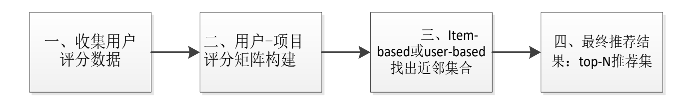


图 2　传统协同过滤算法的一般流程

1.2　基于 slope one 算法改进评分矩阵填充的模型设计

基于上面的内容，本文改进算法的模型如图 3 所示。

1.2.1 模型描述

该模型主要分为两大模块：1）slope-one 算法改进评分矩阵

填充模块：首先利用 slope-one 算法在原始用户-项目评分矩阵中计算用户 u 对项目 i 的可能评分值，区别于其他空值填补法(均值、中值等)填补的单一和可靠性，该算法计算出的预测值从 1-5 都有可能；接着用该评分预测值一一回填原始稀疏的用户-评分矩阵得到处理后的用户-项目矩阵。2）基于用户的协同过滤算法推荐模块：在填充后的用户-项目评分矩阵上，本文用公式（3）计算用户间的相似度，相似度排序得到前 n 个邻居构成的近邻集 S；接着 对 S 中的所有项目对目标用户的推荐度排序，得到前 n 个项目组成的推荐结果集合。

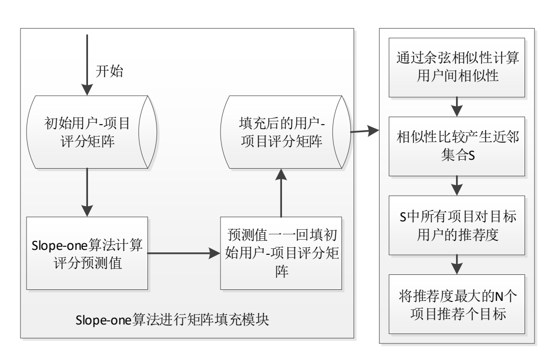


图 3　slope-one 算法改进矩阵填充算法模型

1.2.2 算法流程

基于上面相关概念及算法模型，对本文改进算法的算法流程描述如下所示：

slope-one 算法

1、根据输入的用户评分和项目信息，初始化得稀疏的“用户-项目”评分矩阵

2、for each　in　：(对共同评价过项目 i/j 的用户集合 中的每位用户 u)　用式(1)计算项目 与其他项目的评分偏差

3、for each　in　：(对每个用户 未评价的项目集合 :中的每个) 用式(2)计算用户 对项目 的评分预测值

4、预测值回填初始“用户-项目”评分矩阵，得填充后的“用户-项目”评分矩阵

5、根据填充后的“用户-项目”评分矩阵 用公式(4)计算用户间相似性

6、相似性排序，得到前 k 个用户组成近邻集 S

7、近邻集 S 的所有项目的推荐度计算并排序，采用 top-N 得推荐结果集

8、对测试集中的项目，做评分预测

2　仿真实验与结果分析

2.1　实验数据

实验的数据集使用的是由美国明尼苏达州立大学的 GrouLens 研究小组发布的 MovieLens 标准数据集[18]。本文实验使用的是 1M 的数据集，其中包括 6040 个独立用户对 3952 部电影作品的 1,000,209(100 万余)条评分数据,数据稀疏程度为 1-1000209/(6040*3952)=0.9581。该数据集包括了三个数据文件：存储用户基本信息的 users.dat，存储电影基本信息的 movies.dat 和存储用户对电影评分信息的 ratings.dat。本文实验测试用到后两个数据即 movies.dat 和 ratings.dat。movies.dat 文件中每一条数据表示一部电影信息，其格式为 MovieID::Title::Genres；ratings.dat 文件中每一条数据表示一个“用户-电影”评分信息，其格式为 UserID::MovieID::Rating::Timestamp，其中 rating 范围是 1~5 的整数，分数越高，表示该用户对该电影的喜好程度越大。

2.2　评测指标

评测推荐系统质量的主要度量标准有统计准确度和决策支持准确度。常见统计准确度指标有用户满意度、预测准确度、多样性、惊喜度、新颖性、实时性、健壮性、信任度、商业目标。目前在对离线实验进行评价的时候主要采用预测准确度、Top N 推荐和覆盖率[19,20,21]，本文采用采用预测准确度和 top-N 推荐两个指标对改进后算法进行评测。

a)预测准确度一般通过均方根误差(RMSE)和平均绝对误差(MAE)计算，本文选取后者(MAE)公式如式(4)所示。

$$MAE = \frac{\sum_{u,i \in T} \left| r_{ui} - \overline{r_{ui}} \right|}{|T|} \tag{4}$$

其中：r_{ui} 是用户 u 对项目 i 的实际评分，$\overline{r_{ui}}$ 为算法给出的预测评分，T 为测试集内所有项目集，$|T|$ 为该项目集的大小。

b）Top_N:一般通过准确率(precision)、召回率(recall)和综合值 F 度量，公式如式(5)~(7)所示。

$$\Pr ecision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \tag{5}$$

$$\operatorname{Re} call = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \tag{6}$$

$$F = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{7}$$

其中：$R(u)$ 是用户在训练集上的行为给用户给出的推荐列表，而 $T(u)$ 是用户在测试集上的行为列表。

2.3　实验设计与结果分析

实验采用 Python 程序设计仿真环境，将 80%的评分数据集 rating 作为训练集用于训练模型，剩余 20%作为测试集。，在实验过程中采用 5 折交叉实验验证以避免随机性、偶然性对实验结果的影响。即在 Python 程序中设置 5 个不同随机种子，将 rating 数据集分成 5 份互不相交的子集，每次选取其中 1 份作为测试数据集进行测试，剩余 4 份作为训练集，进行 5 次即每份子集依次作为测试集。

图 4 展示了 UBCF 算法、基于均值填充的协同过滤算法以及本文改进算法的预测准确度 MAE 随邻居数变化的趋势。从图可以看出均值填充矩阵算法同 UBCF 算法的平均绝对偏差都比较大；而经由 slope-one 改进矩阵填充的算法的平均绝对偏差 MAE 较小，并且均值改进算法和 UBCF 算法的结果差别不是很显著，说明均值填充矩阵算法效果不是很明显，而本文改进

后的算法的评分预测准确度相比于协同过滤算法和均值改进算法都有较为显著的提高。另外，当 $K \prec 80$ 时，本文改进算法的 MAE 随着参数 K 的不断增大而减小；当 $K \geq 80$ 时，该 MAE 值趋于平稳。

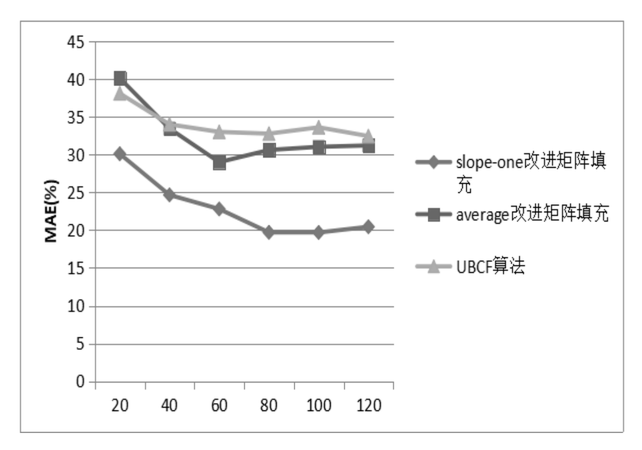


图 4　三种算法平均绝对偏差的折线图比较

为了充分证明 slope-one 算法改进评分填充矩阵的协同过滤算法的推荐准确性，实验还采用了 precision 和 recall 以及综合值 F 三种分类准确度评价指标来作比较。图 5、图 6、图 7 分别描述了随着推荐项目数的变化各算法的 precision 和 recall 以及综合值 F 的变化趋势。

从图 5 三种算法的准确率折线图可以看出，三种算法的准确率变化趋势都是先平稳后上升再有所下降，其中 slope-one 算法改进矩阵填充算法的准确率最好，均值改进矩阵填充的算法的准确率居中，说明两者对协同过滤算法都有一定的改进效果，而本文提出的算法改进效果更加明显。从图 6 三种算法的召回率折现图可以看出，三种算法的召回率变化趋势则为先平稳后下降再有所上升。召回率和准确率变化趋势相反的原因分析：由于在个性化推荐中，召回率和准确率往往是相悖的，通常准确率的提高都是以牺牲召回率为代价的。结合图 5、图 6 的曲线图可以看出本文改进算法使得推荐效果有了更显著的提高，从而也能说明本文提出的改进算法有良好的推荐效果。而图 7 也进一步证实了本文提出的进算法的可行性，确实在一定程度上改善了推荐系统的推荐效果。

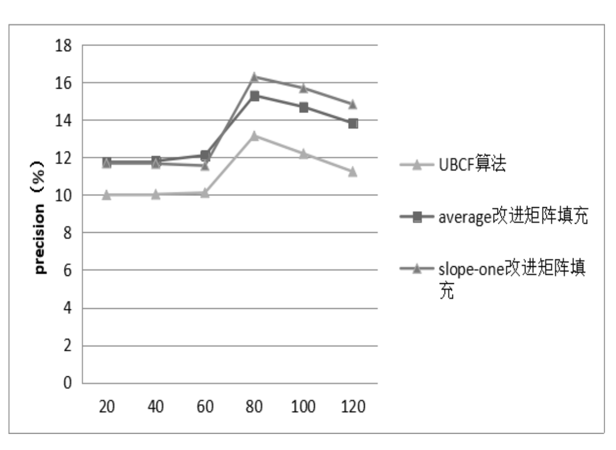


图 5　三种算法准确率的折线图比较

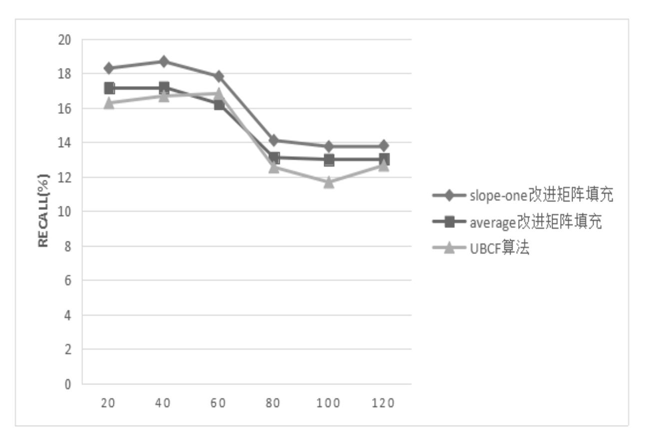


图 6　三种算法召回率的折现图比较

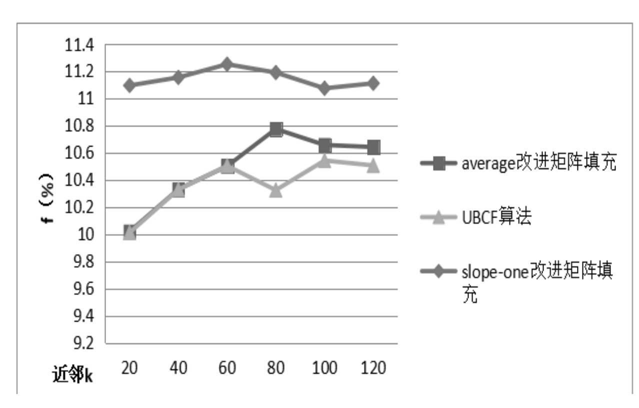


图 7　三种算法准确率召回率的综合指标 f 值的折线图比较

3　结束语

针对传统协同过滤算法的评分数据稀疏和空值填补法过于单一的问题，本文提出基于 Slope One 算法改进评分填充矩阵，先从初始用户-项目评分矩阵出发由 Slope One 算法计算评分预测值后再对稀疏矩阵进行填充，在一定程度上缓解了协同过滤推荐中的数据稀疏性问题，并弥补了空值填补法填补值过于单一的问题。随后，在填充后的用户-项目评分矩阵下给出推荐列表及测试集的预测值，并与其他协同过滤推荐算法进行比较，结果表明本文算法可以改善数据稀疏性问题，并推高推荐系统的推荐质量。

本文的局限在于只是简单的使用了 Slope One 算法进行预测并回填，没有结合其他方法深入的对预测值进行进一步的筛选，下一步的工作准备针对该问题进行深入研究和完善。

参考文献：

[1] 项亮. 推荐系统实践 [M]. 北京：人民邮电出版社, 2012.

[2] 弗朗西斯科·里奇, 等. 推荐系统：技术、评估及高效算法 [M]. 李艳民等, 译. 北京：机械工业出版社, 2015.

[3] 冷亚军, 陆青, 梁昌勇. 协同过滤推荐技术综述 [J]. 模式识别与人工智能, 2014, 27 (8): 720-734.

[4] 孟祥武, 刘树栋, 张玉洁等. 社会化推荐系统研究 [J]. 软件学报, 2015, 26 (6): 1356-1372.

[5] 孔欣欣，苏本昌，王宏志，等. 基于标签权重评分的推荐模型及算法研究 [J]. 计算机学报, 2017, 40 (6): 1440-1452.

[6] Li Liang, Dong Yuxin, Zhao Chunhui, *et al.* Collaborative filtering recommendation algorithm combined with user trust [J]. Journal of Chinese Computer Systems, 2017, 38 (5): 951-955

[7] Gong S. An efficient collaborative recommendation algorithm based on item clustering [C]// Advances in Wireless Networks and Information Systems. 2010: 381-387.

[8] Yan L U, Liu Y. A collaborative filtering algorithm based on predicting and filling missing-data by iterated [J]. Computer & Digital Engineering, 2016, (6): 13-17.

[9] Vozalis M G, Margaritis K G. Applying SVD on item-based filtering [C]// Proc of International Conference on Intelligent Systems Design and Applications. Washington DC: IEEE Computer Society, 2005: 464-469.

[10] Yang B, Lei Y, Liu J, *et al.* Social collaborative filtering by trust [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39 (8): 1633-1647.

[11] 刘业峰，柴天佑. 一种改进的 Slope One 协同过滤推荐算法 [J]. 控制工程, 2017, 24 (2): 257-262.

[12] 李剑锋，秦拯. 一种基于局部近邻 Slope One 协同过滤推荐算法 [J]. 计算机工程与科学, 2017, 39 (7): 1346-1351.

[13] 柴华，刘建毅. 一种改进的 Slope One 推荐算法研究 [J]. 信息网络安全, 2015 (2): 77-81.

[14] 董丽，邢春晓，王克宏. 基于不同数据集的协作过滤算法评测 [J]. 清华大学学报：自然科学版, 2009 (4): 590-594.

[15] Goldberg D. Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35 (12): 61-70.

[16] 刘青文. 基于协同过滤的推荐算法研究 [D]. 合肥：中国科学技术大学, 2013.

[17] 李斌. 推荐系统研究综述 [J]. 现代计算机, 2014 (2): 7-10.

[18] Grouplens. MovieLens data [DB/OL]. http://grouplens. org. University of Minnesota.

[19] 张海朋. 基于协同过滤的电影推荐系统的构建 [D]. 西安：西安电子科技大学, 2015.

[20] Pazzani M, Billsus D. Learning and revising user profiles: the identification of interesting Web sites [M]. [S. l.] : Kluwer Academic Publishers, 1997.

[21] Sarwar B, Karypis G, Konstan J, *et al.* Analysis of recommendation algorithms for e-commerce [C]// Proc of ACM Conference on Electronic Commerce. 2000: 158-167.